Verifikation einer Verbressource anhand der Tiger Treebank

Klenner, M

Verifikation einer Verbressource anhand der Tiger Treebank

Manfred Klenner
Institut für Computerlinguistik
Universität Zürich
klenner@cl.unizh.ch

Abstract

Es wird die automatische Extraktion von Subkategorisierungsrahmen aus einem maschinenlesbaren Verblexikon für Deutschlernende (Deutsch als Fremdsprache) vorgestellt. Die Evaluierung der gewonnenen Verbrahmen erfolgt automatisch im Abgleich mit der Tiger-Baumbank.

1 Einführung

Subkategorisierungsrahmen spielen in der Computerlinguistik eine wichtige Rolle, sei es in formalen Grammatiktheorien wie der LFG oder beim Parsing. Verschiedene Anstrengungen wurden unternommen, Subkategorisierungsrahmen automatisch aus annotierten bzw. nicht-annotierten Corpora (Schulte im Walde, 2002) (Spranger et al., 2006) zu extrahieren. Die vorliegende Arbeit beschreibt ein komplementäres Projekt, bei dem, ausgehend von einer manuell erstellten, idiosynkratisch kodierten und mit "künstlichen", oft auch antiquierten Beispielsätzen[1] illustrierten Verbdatenbank für Deutschlerner, ein für computerlinguistische Zwecke einsetzbares, partiell validiertes Verblexikon mit Subkategorisierungsrahmen generiert wurde. Die Subkategorisierungsrahmen wurden im Stile der LFG mit grammatischen Funktionen spezifiziert. Die verwendete lexikalische Ressource (von nun an: Griesbachlexikon) enthält außer syntaktischer auch semantische Information, die in einem später Schritt validiert werden soll, u.a. mit GermaNet.

Wir beschreiben zunächst die Struktur des Griesbachlexikons.

2 Griesbachlexikon

Das Lexikon von (Griesbach and Uhlig, 2001) erfaßt "rund 90.000 Sprachbelege über den Gebrauch von Verben. (..) Alle Sprachbelege sind aufsteigend codiert nach Satzstrukturen und Valenzkonstellationen. (..) Es gibt genaue Informationen über den Bedeutungsumfang der Verben im Prädikat (Semantik). Es zeigt, mit welchen Worttypen (Kollokation) das Verb im Prädikat zur Beschreibung eines Sachverhalts zusammenwirkt."(vgl. Webseite des Erstautors: http://www.griesbach-daf.de/Verb-Wb.htm).

Das Wörterbuch bzw. die Verbdatei enthält eine alphabetisch geordnete Liste von Einträgen:

- das Verb ohne erweiterten Wortstamm (z. B. "treiben"),
- das Verb mit Präfixen (z. B. "betreiben"),
- das Verb mit festen (untrennbaren) Verbzusätzen (z. B. "untertreiben"),
- das Verb mit unfesten (trennbaren) Verbzusätzen (z. B. "ab-treiben").

Es werden (zumindest teilweise vom Autor selbst erzeugte) Beispielsätze gegeben, die zur Illustration der Kodierung dienen. Diese erfolgt anhand von zwei Codeziffern: *Ss* für die Satzstruktur (u.a. Selektionsrestriktionen) und *Fk* für die sog. Funktionskennzeichen (u.a. Kasusinformation).

Die (in der Regel 4-stellige) Kodierung der Satzstrukturen (Ss) und auch der Funktionskennzeichen (Fk) spezifiziert die (wichtigsten) grammatischen Funktionen des Verbs:

- Die erste Stelle steht für das Subjekt.
- Die zweite Stelle steht für das direkte Objekt (1. Objekt).
- Die dritte Stelle steht für das indirekte Objekt (2. Objekt).
- Die vierte Stelle steht für präpositionale Ausdrücke

[1] Zum Teil auch mit nicht mehr gebräuchlichen Anredeformen wie "Fräulein", was im DaF-Kontext besser zu vermeiden wäre.

Jede Stelle (Pos, vgl. folgende Tabelle) wird mit Ziffern zwischen 1 - 9 belegt. Wir betrachten zuerst die Kodierung für die Satzstrukturen (Ss).

Pos	Funktion	Kodierung/Bedeutung
1.	Subjekt	1/Person, 2/Sache, 3/Begriff 4/SV, 6/Präd.subj., 9/'es'
2.	1. Objekt	1/Person, 2/Sache, 3/Begriff 4/SV, 6/Präd.obj., 8/'sich', 9/'es'
3.	2. Objekt	1/Person, 2/Sache, 4/SV, 8/'sich'
4.	PO etc.	1/lokal, 2/temporal, 4/kausal 5/final, 6/Präd.nom/-akk 7/Präd.dat/-gen, 8/austauschbare und 9/feste Präd.ergänzung

PO steht für Prädikatsobjekt, SV meint Sachverhalt.

Die Kodierung der Funktionskennzeichnung (Fk) jeder dieser Positionen dient vor allem der Identifikation der Kasusinformation. Hier wird aber auch Information über Präpositionalobjekte (welchen Kasus hat die Präposition) abgelegt und Information, ob das Verb Satzkomplemente nimmt:

1	Nominativzeichen	5	Akkusativpräposition
2	Akkusativzeichen	6	Dativpräposition
3	Dativzeichen	7	Akk-/Dat-Präposition
4	Genitivzeichen	8	Genitivpräposition
		9	Konjunktion

Zudem gibt es Kürzel für die Aktionsarten: H (= Handlung), V (= Vorgang), S (= Sein/Zustand). Die Autoren geben folgende Erklärung: (Handlung:) Peter kommt ins Krankenhaus (, um mich zu besuchen). (Vorgang:) Peter kommt ins Krankenhaus (, weil er krank/verletzt ist). (Sein:) Peter ist krank/verletzt.

Zum Abschluss dieses Überblicks sei ein Beispiel gegeben, das Verb "auswachsen"

- aus-wachsen (abtrennbares Präfix)
- hat (Perfekthilfsverb)
- "die Unruhe im Volk wächst sich aus" (Bsp.)
- V (Aktionsart: Vorgang)
- Ss 4800 (Subjekt ist ein Sachverhalt, 1. Objekt ist "sich")
- Fk 1200 (Subjekt ist Nominativ, 1. Objekt Akkusativ)

Das Lexikon umfasst sehr viele heterogene Daten, die recht idiosynkratisch kodiert sind. Ihre Verlässlichkeit ist ohne eine automatische Verifikation praktisch nicht einzuschätzen. Ein Vergleich mit empirischen Ressourcen wie automatisch extrahierte Verbrahmen, oder Baumbanken kann allerdings Abhilfe schaffen. Für die "semantische" Information ist GermaNet ein möglicher Verifikationshintergrund. Es könnte sein, dass das Lexikon eine reichhaltige Ressource ist, die für computerlinguistische Zwecke geeignet ist - zumindest nach einer Transformation und anschließenden Verifikation. Dieses herauszufinden war Ziel der vorliegenden Arbeit.

3 Abbildung Griesbach-Tigerlabel

Die Extraktion der Griesbach-Verbrahmen und ihre Abbildung auf Tigerlabel ist weitgehend trivial. Eine Kombination aus Angaben in Ss bzw. Fk ist meist ausreichend und eindeutig. Hier die wichtigsten Abbildungen:

Griesbach	Tigerlabel
Subjekt (Ss) & Nom. (Fk)	SB
1. Objekt (Ss) & Akk. (Fk)	OA bzw. OA2
1. Objekt Code 9 (Ss)	EP ("es")
2. Objekt (Ss) & Dativ (Fk)	DA
Genitivzeichen bzw. phrasaler Genitiv	OG bzw. PG
präpositionales Objekt (Fk)	OP evtl. CVC
Konj (Fk)	OC

4 Extraktion von Verbrahmen aus der Tiger-Baumbank

Die Tiger-Baumbank (Brants et al., 2002) ist ein mittlerweile auf 50000 Sätze (Release 2) angewachsenes, syntaktisch annotiertes Korpus des Deutschen. Die Knoten sind syntaktische Label wie NP etc., die Kantenbeschriftungen markieren die grammatische Funktion dieser Phrasen. Z.B. markiert NP-SB eine Subjektnominalphrase. Das Tiger Searchtool kann dazu benutzt werden, eine empirische Exploration und Auswertung der Daten zu unternehmen. Es besitzt eine graphische Ausgabe, eine eigene Suchsprache und die Möglichkeit Statistiken zu erstellen (und in XML zu exportieren). Leider kann die Tigersuchsprache nur manuell im Searchtool verwendet werden, zur automatischen Extraktion und vor allem automatischen Weiterverarbeitung der Resultate kann es nicht dienen. Zwar kann man z.B. bestimmte Knoten mittels Variablen "binden", aber es fehlt die Möglichkeit, die gesammelte Information neu zusammengestellt und in einem benutzerdefinierten Format auszugeben (es gibt nur das wenig flexible, trotzdem aber nützliche Tabellenformat).

In (Klenner et al., 2004) wird ein Interpreter zur automatischen semantischen Annotation aus Baumbanken beschrieben, der bis auf einige Ausnahmen syntaktisch der Tigerabfragesprache nachempfunden ist. Dieser Interpreter ist in Prolog geschrieben und er operiert über Baumstrukturen, die ebenfalls in Prolog repräsentiert sind, in einem sehr einfach zu verarbeitenden Format. Es gibt 3 Prädikate, eines für Dominanz (id), eines für Präzedenz (lp) und eines für kategorielle bzw. funktionale Merkmale (feature/3), etwa die Wortform, das syntaktische Label, oder die STTS-Wortklasse.

- id(Satznummer,Mutterknoten,Tochterknoten).
- lp(Satznummer,Knoten,Nachfolgerknoten).
- feature(Satznummer,Knoten,Feature).

Die Knoten sind Ziffern wie sie z.B. in Negra und Tiger verwendet werden: 1-499 für Terminale und ab 500 für Nicht-Terminale. Über der Prologvariante der Baumbank lassen sich sehr leicht Suchprädikate formulieren, mittels Operatordefinitionen wurde die Syntax von Tiger nachempfunden. Hinzukommt ein Aktionsteil als Ergänzung. Dieser dient dazu, beliebige Aktionen mit den gefundenen Bindungen auszuführen.

Hier ein Beispiel einer Extraktionsregel:

```
#z >SB #x  &
#z >HD #y
==>
frame(sb,#x,#y).
```

#z, #x und #y sind lokale Variablen, die durch den Interpreter an Knotennummern im Baum gebunden werden. Es gilt implizit, dass #x $\neq$ #y $\neq$ #z. Das Prädikat "frame" dient hier der Bestimmung einer Kasusrolle des Verbs. So wird im vorliegenden Fall ein Fakt assertiert, das die Satznummer, die Verbnummer, und die Nummer der NP als Subjektrelation ablegt. In einem nachgeschalteten Schritt werden alle so assertierten Relationen eines Verbs zu dessen Kasusrahmen zusammengefasst.

Komplexere Regeln, bei denen auf einen Streich ein kompletter Kasusrahmen abgegriffen wird, sind denkbar. Nur praktische Erwägungen sprechen dagegen - solche vollständigen Pattern erfordern aufgrund der möglichen Permutationen grammatischer Funktionen viel zu viel Regeln.

Ursprünglich diente der Interpreter zur semantischen Interpretation im Kontext von Antwortextraktionssystemen, etwa zur Generierung minimaler logischer Formen, vgl. (Hess, 1998). Er kann mit kleinen Änderungen aber für beliebige Extraktionsaufgaben (z.B. Vektorgenerierung) verwendet werden.

Die Resultate der Extraktion von Kasusrahmen sind von hoher Präzision, da die Extraktionsregeln über einer manuell annotierten Baumbank operieren und die Strukturen der Baumbank eindeutig sind. Es kam dabei nicht auf Vollständigkeit an, z.B. wurden Kasusrahmen in Verbkoordinationen nicht erfasst.

5 Empirische Ergebnisse[2]

Die von uns extrahierte Version des Griesbachlexikons weist 10033 Verblemmata auf, wobei davon 1084 Verbwurzeln bilden, der Rest also aus Partikelkonstruktionen hervorgegangen ist. Erste Stichproben zeigten, dass das Griesbachlexikon weit entfernt von einer umfassenden Abdeckung der deutschen Verben ist. Verben wie "stagnieren, solidarisieren, verbessern, gähnen, erobern, beten", die durchaus keine Spezialfälle darstellen, fehlen.

Es wurden 5803 Verbrahmen aus Tiger extrahiert. Davon sind 1200 nicht im Griesbach (GB)[3], was 712 Verben entspricht. Vollständige Übereinstimmung der Kasusrahmen wurde bei 3231 Rahmen erzielt, 1270 überlappen lediglich. Entweder inkludiert der Griesbachrahmen den Tigerrahmen oder umgekehrt. In 102 Fällen war keine Übereinstimmung vorhanden.

	Anzahl
Verbrahmen nicht in GB	1200
Rahmenübereinstimmung	3231 (validiert)
Tiger/ GB hat mehr	1270
keine Übereinstimmung	102

Alle die Verbrahmen, die übereinstimmen, können als validiert gelten, hier 3231 Stück. Alle die Verben, die im Tiger und im Griesbach vorkommen, aber in den Kasusrahmen nicht übereinstimmen, sind Kandidaten für eine manuelle Fehlerkontrolle. Insbesondere auch dort, wo ein Griesbachkasusrahmen für ein Verb keine Entsprechung in Tiger findet, liegt mit ziemlicher Sicherheit ein Fehler vor. Aus diese Weise haben wir fehlerhafte oder zumindest zweifelhafte Kodierungen im Griesbach aufgedeckt:

- 'Kasuskodierung falsch": z.B. 1400-1200

[2] Vgl. auch (Luethi, 2005) - eine Seminararbeit, die die vorliegenden Ergebnisse qualitativ bestätigt.

[3] Elliptische Sätze im Tigerkorpus führen zu defizitären Rahmen und verzerren diese Statistik leicht.

"er weiß zu schweigen, wenn es sein muss"

"zu schweigen" als Akkusativ (die 2 in 1200)

- "Kodierung fehlt": z.B. 1400-1900

  "sie sagte ihm, dass er gehen könne"

  "ihm" ist 3. Stelle, bleibt unkodiert

- "Kodierung einer Wortart falsch": 1280-1239

  " er hat sich einen Sessel herbeigezogen"

  die 9 in 1239 postuliert eine (nicht vorhandene) Konjunktion

Fehler, die nicht automatisch identifizierbar, aber beim Einlernen in die Notation entdeckt wurden (man darf vermuten, dass es noch mehr davon gibt): Das Verb "fortsetzen" hat folgenden Beispielsatz: 'die Verhandlungen werden nächste Woche fortgesetzt'. Der Satzstrukturcode ist 3800, was bedeutet, dass das Subjekt ein Begriff ist (Verhandlungen) und das 1. Objekt durch das Reflexivpronomen 'sich" realisiert sein müsste. Der Satz allerdings enthält kein Reflexivpronomen.

Diese Fehler zeigen die Anfälligkeit (nicht den heutigen Computerlinguistikstandards entsprechender) manuell erzeugter Ressourcen auf und legen eine verstärkten Einsatz von CL-Techniken (in Verbindung mit CL-Ressourcen) zu deren Evaluierung nahe. Die umgekehrte Hoffnung, Nicht-CL-Ressourcen CL-tauglich zu machen, ist damit nicht gestorben, aber wie sich zeigt nicht einfach automatisierbar. Von den vielen partiellen Verbrahmenressourcen wie Celex (Baayen et al, 1993), Telex (Kunze, 1991) hat Griesbach den Vorteil, auch Selektionsrestriktionen anzugeben. Dieser Teil bleibt zu evaluieren.

6 Zusammenfassung

Die validierten Rahmen wurden in (Klenner, 2005) und (Klenner, 2006) zur Auszeichnung von Chunks und Verben hinsichtlich grammatischer Funktionen verwendet. Eine andere Verwendung unseres Ansatzes könnte die automatische Selektion realer, d.i. aus einem Korpus stammender Beispielsätze für die Griesbachverben sein - was die Qualität der Ressource für den Deutschunterricht verbessern würde. Außerdem haben wir automatisch eine Reihe von Fehlern entdeckt, die zur Bereinigung der Orignalressource beitragen können (Rückmeldung an den Autor). Ziel unserer Arbeit ist aber in erster Linie eine verifiziertes Subkategorisierungslexikon für deutsche Verben. Inwieweit die "semantische" Information aus dem Griesbach-Lexikon brauchbar ist, soll im nächsten Schritt getestet werden.

Danksagung: Dank an Heinz Griesbach für die maschinenlesbaren Variante seines Wörterbuchs. Mein Dank gilt auch Martin Volk, der die initialen Konvertierungsschritte unternommen hatte und Simon Clematide für die Konzeption der Prologdarstellung von Syntaxbäumen.

Literatur

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius and George Smith. The TIGER Treebank. *Proc. of the Workshop on Treebanks and Linguistic Theories Sozopol*. 2002.

R.H. Baayen, R. Piepenbrock and H. van Rijn. The CELEX Lexical Database (CD-ROM). Linguistic Data Consortium. Univ. of Pennsylvania. 1993.

Heinz Griesbach und Gudrun Uhlig. Großes Verb-Wörterbuch. *BoD GmBH, Noderrstedt*. 2001.

Michael Hess. Antwortextraktion über beschränkten Bereichen. *Proc. of KONVENS. Bonn,* 1998.

Manfred Klenner, Fabio Rinaldi and Michael Hess. Steps towards Semantically Annotated Language Resources. *Proc. of LREC*. 2004.

Manfred Klenner. Extracting Predicate Structures from Parse Trees. *Proc. of the Intern. Conf. on Recent Advances in NLP (RANLP)*. 2005.

Manfred Klenner. Grammatical Role Labeling with Integer Linear Programming. *Proc. of EACL, Conference Companion, pp. 187-190*. 2006.

Jürgen Kunze. Kasusrelationen und Semantische Emphase. (Studia grammatica XXXII). Berlin: Akademie Verlag. 1991.

Katrin Luethi. Evaluation des Verblexikons von Griesbach anhand des TIGER-Korpus. Seminararbeit, Univ. Zürich. 2005.

Sabine Schulte im Walde. Evaluating Verb Subcategorisation Frames learned by a German Statistical Grammar against Manual Definitions in the Duden Dictionary. *Proc. of the 10th EURALEX Intern. Congress*. 2002.

Sabine Schulte im Walde. Induction of Semantic Classes for German Verbs. *Stefan Langer und Daniel Schnorbusch (eds). Semantik im Lexikon. Gunter Narr Verlag, Tübingen*. 2004.

Kristina Spranger, Martin Forst, Margit Gut, Ulrich Heid, Hannah Kermes, Christian Rohrer und Melvin Wurster. Verfahren zur effizienten Ergänzung von Sukategorisierungslexika für die maschinelle syntaktische Analyse. *Poster auf der 27. Jahrestagung DGfs* . 2005.